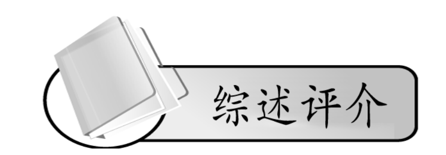


时态信息检索研究综述*

张晓娟　韩　毅
(西南大学计算机与信息科学学院　重庆 400715)

摘要:【目的】总结国内外时态信息检索研究现状，以期为相关学者更好地把握时态信息检索研究问题提供理论基础。【文献范围】在 Google Scholar 中分别以检索式"Temporal Information"与"时态信息"且不限定时间范围地进行文献检索，获得部分相关文献后，再结合追溯法最终得到 92 篇相关文献。【方法】基于文献调研与归纳总结方法，分别从文档中时态信息抽取、查询中时态信息识别和时间感知排序三方面对时态信息检索的相关研究进行综述与评述。【结果】研究发现时态信息检索研究存在着如下问题和挑战：国外对时态检索研究比较多，而国内的相关研究甚少；利用表征时间信息的实体与事件演化信息识别文档关注时间的相关研究不足；缺乏对非周期变化查询的意图预测；时态信息检索模型实验的可重复性有待提高。【局限】未对该领域的文档采集、文档索引以及相关应用进行文献综述。【结论】构建标准化的评测数据集以及无参数时态信息检索模型将是时态信息检索领域的未来方向研究。

关键词: 时态信息检索　时态信息　时态查询意图　时态感知排序
分类号: G350

1 引 言

随着网络中数字资源的迅速增长，每天都有大量新文档生成和旧文档更新，产生了诸如网络存档(Web Archives)、新闻报告、博客与个人邮件等与时间因素有关的数据集。因此，如何从此类数据集中为用户提供及时可靠信息是当前检索系统的首要任务。然而，基于关键词匹配的传统检索模型难以从时间数据集中为用户返回满意的检索结果，其主要原因在于：用户提交的查询与时间相关，如 Metzle 等[1]通过对 AOL 查询日志分析得出，约 1.5%的查询具有显式时间意图(如查询"SIGIR 2016")，约 7%的查询包含隐式时间意图(如查询"北京奥运会")；其次，时间数据集中除存储了最新文档外，也存储了文档在其他时间段的不同形式，故数据集中文档也具有时间属性。而传统信息检索模型常常忽略了查询中时间属性是否与文档时态属性匹配这一特定条件。为解决此问题，Temporal Information Retrieval(T-IR)应运而生，其目标是利用查询与文档中时态信息来提高最终检索准确度。因国内学者[2-5]在从事"Temporal Information"相关研究时，将其普遍翻译为"时态信息"，故笔者在本文中将"Temporal Information Retrieval"译为"时态信息检索"。

时态信息检索在其他相关任务(如文档探索、相似性搜索和信息聚类等)也起着重要作用，此研究引起了信息检索及其相关领域的广泛关注。其中，信息检索领域的一些重要国际会议(如 SIGIR、WWW、CIKM、NTCIR-11 与 NTCIR-12 等)对时态信息检索相关研究给予了高度重视，使其成为近年来网络信息检索领域探讨的热点话题。综合已有研究，时态信息检索的研究领域主要包括：文档采集、文档索引、文档与查询

通讯作者：张晓娟，ORCID：0000-0002-5889-5922，E-mail：zhangxiaojuan624@gmail.com。

*本文系国家社会科学基金青年项目"融合用户个性化与实时性意图的查询推荐模型研究"(项目编号：15CTQ019)和西南大学博士启动基金项目"查询意图自动分类与分析研究"(项目编号：SWU114093)的研究成果之一。

中时态信息抽取，时态感知排序以及 T-IR 的相关应用如时态文档的自动摘要、聚类与自动分类等[6-7]。由于时间与精力有限，笔者难以对其所有相关研究进行文献综述，因此只是基于传统信息检索的角度，从文档中时态信息抽取、查询中时态信息识别与时态感知排序三个方面对国内外时态信息检索的相关研究进展进行总结与评述。

笔者在 Google Scholar 中分别以检索式“Temporal Information”与“时态信息”且不限定时间范围地进行文献检索，再根据研究主题进行筛选后得到部分相关文献。然后，在这些相关文献基础上，进一步利用追溯法最终共获得 92 篇文献。其中，英文文献 86 篇，而中文文献 6 篇。本文尝试通过较全面的文献调研对时态信息检索这一研究课题的国内外研究进展进行较为系统的分析评述，以期为相关学者更好地把握时态信息检索研究问题提供理论基础。

2 文档中时态信息抽取

文档中主要包括以下四类时态信息[8]:

①日期，表示能在日历上查找到的时间表达式，如“2016 年 7 月 23 日”、“上周一”等;

②时间，表示一天中某个具体或者模糊时间段，如“18 点 20 分”、“中午”、“1 月 2 日的上午”等;

③时间区间，表示某个具体时间段，如“24 个月”、“从 2013 年至 2017 年”等;

④时间集合，如“每隔两周”、“一周两次”等。

其中，文档中时态信息抽取主要包括文档时态元数据抽取与文档关注时态信息两方面[7]。

2.1 文档时态元数据抽取

文档时态元数据主要包括文档的创建时态信息、采集时态信息和最新修改时态信息；文档的采集时态与最新修改时态信息常保留在网络服务器中，可直接获取[9]。因此，文档创建时态信息抽取是时态元数据抽取研究的主要内容，主要有基于内容和链接结构的两类抽取方法。

(1) 基于内容的文档创建时态信息抽取主要是借助文档中词信息来识别其创建时间，具体实现思想为：首先对文档创建时间的可能时间段进行预定义，再根据这些时间段对数据集进行分类，最后利用数学模型计算文档中词在各个时间段文档集合中的出现情况，以此来确定该文档的创建时间。其中，常用的数据模型为时间语言模型[10]和时间熵[11]。除此之外，一些学者也尝试采用其他方法，如 Chambers[12]提出从文档所包含的时态表达式中抽取特征来训练分类器以识别文档创建时间，其实验结果表明该方法的效果优于基于数学模型的方法，但缺陷在于只能识别该文档创建于哪一年，无法对更加细粒度的时间信息加以识别；Kotsakos 等[13]在假设内容相似文档之间创建时间相近的基础上，尝试在不给定时间粒度的情况下对文档创建时间信息进行识别，即利用统计学方法分别计算两文档之间重要词的突发区间，再将这些突发区间之间的重叠区域视为文档的创建时间段；Garcia-Fernandez 等[14]借助外部知识资源(如 Google Book N-gram、Wikipedia 以及词源学的背景知识)，采用监督式与非监督式方法来识别旧法语新闻报纸的出版时间；Tilahun 等[15]利用与文献[10-11]相同的方法识别了中世纪时期拉丁文英国宪法的创建时间。总之，基于内容的方法为当前文档创建时态信息抽取的主流方法，其优点在于简单且易于实现，而缺点是最终识别的准确度依赖于对数据集时间范围划分的准确度，且可识别的文档创建时间范围取决于数据集所包含的时间范围。另外，以上研究都假设每个文档都只存在着一个特定的创建时间，而 Zhao 等[16]认为这种假设只是存在于新闻数据集中，在真实网络环境中，不同文档的不同部分创建时间可能不一样，如博客数据集，其子文档(博客条目)会有不同的创建时间。于是，通过对 ClueWeb 12 数据集①中每个文档的每个段落进行时间标注，实验结果发现，约三分之二的文档中子文档的创建时间不同，故如何进一步对文档中子文档的创建时间识别将是后续研究工作中需探讨的问题。

(2) 基于链接结构的文档创建时态信息抽取方法的主要思想为：首先借助文档之间链接结构构建图模型，再通过相关模型遍历图，最后利用图中与某文档相邻接的文档的创建时态信息来识别该文档的创建时态，如 Nunes 等[9]与 Salah 等[17]构建图模型后，在采用一步传递方法(One-step Propagation)遍历图的基础上，Nunes 等[9]将邻接文档中最后修改时间的平均值作为

①http://lemurproject.org/clueweb12/.

该文档的创建时间, Salah 等[17]将邻接文档中最新创建时间作为此文档的创建时间; Prokhorenkova 等[18]分别采用一步式或多步式(Multi-step Propagation)传播模式遍历文档图模型来识别文档创建时间, 其实验结果表明, 基于多步式传递方法优于一步式传递方法。总之, 基于非文档内容方法的主要优点是不需预先确定文档可能所属的时间范围, 可识别任意时间段文档的创建时间, 但该方法识别的准确度依赖于其他文档时态信息的可获取性与准确性。

2.2　文档关注时态信息抽取

文档关注时态信息是指文档内容所涉及的时间区间, 主要通过识别与排序文档中时态表达式来获得。其中, 时态表达式主要包括显式(Explicit)、隐式(Implicit)与相对(Relative)三类[19]。显式时态表达式表示某一具体时间点, 其时态粒度可以是某年、某月或者具体某日, 如表达式"2015"、"2015 年 10 月"与"2015 年 10 月 1 日"; 隐式时态表达式是一些借用假日或者事件名称表达相关时间信息, 如"Mothers' Day 2016", 此类表达式在时间轴上固定, 需对其进一步标准化为具体时间"2016 年 5 月 8 日"; 相对时态表达式需借助参考时态信息(如文档内容或者文档创建时间)才能得知其所表达的具体时态信息, 如表达式"今天"与"上周一"。时态表达式识别研究在信息抽取相关的国际会议如 SemEval、Message Understanding Conference (MUC)、Automated Content Extraction (ACE)上进行了广泛探讨, 其相关技术与方法比较成熟, 且已有相关的开源工具可供直接使用, 如 TempEx[20]、GUTime①、HeidelTime[21]与 SuTime[22]等。

时态表达式排序的相关研究有: Strötgen 等[23]综合考虑文档、数据集与查询等特征, 为文档中时态表达式进行排序; Jatowt 等[24]把与文档内容相关事件的发生时间段作为该文档的关注时间, 首先利用聚类算法对一些新闻网页进行聚类, 再根据每个类簇内容与文档内容的相似性, 最终将每个类簇中所包含事件的平均时间段作为该文档的关注时间; Jatowt 等[25]为避免时间表达式与文档关注时间之间的误差, 借助新闻语料集合, 采用时间熵与时间峰度两指标衡量文档中词与某新闻语料的相关度, 最后将相关度最大的新闻语料时间作为该文档的关注时间; Jatowt 等[26]首先从新闻数据集抽取与时态有直接关联的词, 再通过统计学方法综合考虑与文档中词相关的时态信息而最终确定该文档的关注时态, 该方法的优点在于能对不包含或只包含少量时态表达式的文档进行关注时间识别; Zhao 等[27]首先对新闻文档中时态表达式进行抽取与归一化处理, 再提出关系模型(Relation Model)构建文档主题与时态表达式之间关系, 以此识别新闻文档中的主题时间; Kumar 等[28]首先在维基百科人物自传数据集中利用监督式语言模型训练词随时间(年份)的分布概率, 再以此识别非人物自传 Wikipedia 网页的关注时间(年份)。以上研究都是基于文档级别, 也有学者探讨如何抽取与词相关的时态信息, 如 Spitz 等[29]基于 Wikipedia 语料集, 根据词与时态表达式(如某天、某月或某年)在文档同一句子中共现情况构建加权二部共现图, 再利用类似协同过滤算法为每个词识别相关时间信息或者为相关时间返回词信息, 该研究以期能为相关研究如文档关注时间抽取和时态文档聚类等提供一定技术基础。

相对于文档元数据抽取来说, 文档关注时态信息抽取过程涉及到对文档内容语义理解过程, 其难度相对较高。另已有文档关注时态信息抽取方法大多只是停留在对三类时态表达式的识别与排序, 忽略了文档中表征时间信息的实体以及事件两类重要因素[30-33], 故如何进一步通过跟踪实体与事件随时间的演化信息来进一步提高文档关注时态信息识别准确度将是未来研究的一个重要趋势。

3　查询中时态信息识别

查询是用户信息需求的简化形式[34], 可能包含与文档中类似的时态信息。通过对查询的处理与分析, 有助于判断查询中是否具有时态意图以及用户随时间变化可能感兴趣的潜在查询子主题。基于此, 本文主要对时态意图识别与查询动态子主题识别相关研究进行综述。

3.1　时态意图识别

时态意图识别旨在判断用户提交某查询后是否想获得特定时间段的信息, 综合已有研究, 主要分为:

①http://www.timeml.org/site/tarsqi/modules/gutime/download.html.

给定类目体系下的时态意图识别与不给定类目体系的时态意图识别。

(1) 给定类目体系下的时态意图归类与识别。首先确定查询中可能包含的时态意图类别，再利用相关方法对查询进行自动归类。根据所依赖的数据集，此类研究又可分为基于日志与文档集两类方法。基于日志方法的主体思想是从查询日志中选取分类特征，以此实现各时态意图类别的区分。相关研究主要有：Vlachos 等[35]将查询可能具有的时态意图归为周期查询、季节性查询与大峰值三类，且首次提出利用突发点识别方法对这三类查询自动归类；Parikh 等[36]基于查询中所包含突发点的形状以及停留时间，利用突发点识别方法对时态查询进行自动识别；Kulkarni 等[37]分别从查询中包含的波峰数、波峰形状、波峰趋势以及周期性对时态查询进行归类；Zhang 等[38]采用机器学习方法，从查询日志中选取特征判断某查询是否与公共事件、公共节日或者电视节目等相关；König 等[39]借助机器学习思想判断查询点击垂直新闻搜索结果的楷率，以此判断该查询结果中是否应该融合新闻网页；Ren 等[40] 将查询中可能包含的时态模式归类为稳定性查询(Stable Queries)、一次性突发查询(One-time Burst Queries)、周期性多次突发查询(Periodic Multitime Burst Queries)与非周期多次突发查询(Aperiodic Multitime Burst Queries)，基于时间序列，根据查询在查询日志中的搜索量选取分类特征，利用SVM 分类训练分类模型以此对 4 类查询自动区分。

基于文档集方法的主体是从网络文档集或者外部知识资源如 Wikipedia 中选取分类特征，以此实现各类别时态意图的有效区分。根据所采用的分类体系，此类研究又可细分以下三方面：

①基于 Jones 的时态意图分类体系。Jones 等[41]根据查询返回文档的时态属性，将其分为时间非歧义性(发生在特定时间)、时间歧义(发生在几个可能的时间段)和时间查询(任意时间)，且利用结果文档集为查询构建时间档案，再采用决策树方法实现三类查询的自动分类；基于该分类体系，Campos 等[42]从网络文档片段的标题、文本内容以及链接信息中选取特征实现时间查询的自动分类。

②基于 NTCIR 会议中所提供的时态意图分类体系。NTCIR-11[43] 与 NTCIR-12[44] 会议中时态获取子任务(Temporalia)将查询中可能包含的潜在时态意图分为当前(Recency: 获得当前事件的相关信息)、过去(Past: 获得过去相关事件信息)、将来(Future: 查询预测或者预订的相关事件)与非时间(Atemporal: 如导航类查询)意图。基于此分类体系的主要相关工作有：Yu 等[45]通过选取时间间隙特征(即查询提交时间与查询关注时间之间的时间差)、词时态特征和命名实体特征，再分别采用半监督式和监督式线性分类器训练分类模型，最终实现时间查询的自动分类，且取得了较好的实验效果；Zhao 等[46] 先利用维基百科中概念(Wikipedia Concepts) 扩展查询中可能包含的概念信息，再利用 Wikipedia 网页浏览日志信息抽取与查询概念相关的时间序列数据，以此为获得查询特征信息并实现各时态意图的自动识别；Pei 等[47]选取显示特征(查询上下文词特征) 、隐式特征(利用 Google Trends 进行时间序列分析而得到的时间间隙特征)以及文本特征(词在不同时态意图类别中的概率分布以及词的时态标记信息)选取分类特征训练分类器；Fernando 等[48]选取查询相关特征，如查询中动词时态特征、查询表达式中时间与查询提交时间之间差值特征以及查询中包含的 n-Gram 词元在每个意图类别中的多项分布特征，再通过基于规则的投票方法融合各类特征以此计算每个查询在每个意图类别中的分布概率。

③基于其他时态意图分类体系。Amodeo 等[49]首先根据查询中包含的时态属性将其分为周期性、部分周期性、基于趋势以及随机 4 类，基于纽约时报(New York Times)数据集，利用综合概率与时间序列的启发式模型对查询进行自动分类，且预测查询主题相关的将来事件；Dong 等[50]通过选取分类特征，训练分类模型识别查询是否与某突发性新闻事件相关；Styski 等[51]利用 30 个特征训练了回归模型分类器来预测某查询是否与最新内容相关；Cheng 等[52]通过分析查询词在相关文档中的分布变化判断查询是否具有时态意图。

(2) 不给定类目体系的时态意图识别。即在不给定时态意图类别体系的情况下，利用相关方法判断某个查询与某特定时间或某事件相关。相关研究主要有：Kanhabua 等[31]在未提供时态类别的情况下，将查询时间文档的创建时间视为该查询的关注时间，分别利用查询关键词、返回结果中排序前 K 的文档内容及其相关时间标记来识别查询中包含的时间信息，其中第一种方法是基于查询关键词的语言模型，后两种方法均是基于伪相关反馈思想；Kanhabua 等[53]从查询日志与外部数据集中选取分类特征，利用机器学习方法判断查询是否与某事件相关；Campos 等[42]分别利用查询结果片段与 Google 和 Yahoo 查询日志识别隐式查询内容的关注时间，其实验结果表明利用查询结果片段的方法优于利用查询日志的方法；Zhang 等[38]综合从查询日志与查询结果中选取特征训练分类器识别查询是否与某周期性发生的事件相关；Nguyen 等[54]借助网络文

档中的锚文本数据识别查询子主题所包含子主题的日期信息。

总体来说，当前大多数研究者针对给定类目体系下的时态意图识别的研究多于不给定类目体系的时态意图识别研究。其中，在不给定类目体系下的时间意图识别研究中，基于查询日志的方法有助于通过识别具有相似时间模式查询的时态意图，但该方法在大多数情况下只能对高频查询进行有效识别，而对于低频查询存在着数据稀疏问题；基于文档集的方法能解决数据稀疏的问题，但容易产生一些噪声数据，影响最终识别的准确度。整体来说，基于查询日志与基于文档集方法各有优缺点，但在大多数情况需采用两种方法结合的方式[38, 42]。在不给给类目体系下的时态意图识别研究中，其最终识别的准确度依赖于对相关时间或者事件识别的准确度。

3.2 查询动态子主题识别

对于歧义性时态查询来说，用户在不同时间段所感兴趣的子主题可能不一样，如查询"汶川"，用户可能感兴趣的是与汶川相关的人文地理(汶川地震之前)或汶川地震相关新闻(汶川地震之后)。因此，准确识别用户不同时段对此类查询可能感兴趣的主题显得尤为重要。根据其所依赖的数据集，其研究分为基于查询日志与基于文档的查询动态子主题识别。

(1) 基于查询日志方法的研究内容

①不同时间段的查询子主题识别。在不同时间段根据点击信息、查询之间的语义相似度构建 query-url 二部图，通过相关遍历算法(如随机游走)为每个查询构建向量，再借助聚类算法对查询进行自动聚类，最后将每个类簇质心作为该查询的一个子主题[55-56]。

②时态意图变化趋势预测与建模。利用相关方法识别查询周期性的变化规律获取该查询的将来时态意图：如 Metzler 等[1] 根据查询词与时间限定词在查询日志中的共现识别周期性查询；Shokouhi[57]根据查询日志中历史频率分布，利用时间序列方法判断该查询是否是周期性查询；Radinsky 等[58]基于用户历史行为数据，提出一种 DML 学习算法(Dynamics Model Learner)识别与预测用户意图变化的趋势、周期性及噪音。

(2) 基于文档集方法的研究内容

①借助查询在不同时间段返回的文档集来挖掘其可能的潜在子主题，如 Nguyen 等[55]尝试利用 LDA(Latent Dirichlet Allocation)模型从查询相关文档中进行潜在主题分析，以此识别该查询的动态子主题；Gupta 等[59-60]首先利用一元检索模型为每个查询返回排名前 K 的文档构建伪相关文档集，再利用伪相关文档的出版日期与文档内容中的日期表达式构建生成模型识别该查询不同粒度(如年、月、日)的时间段中用户所感兴趣的内容；Dakka 等[61]提出利用文档的发布时间识别隐式查询可能感兴趣信息的时间段。

②通过挖掘 Wikipedia 层级结构识别查询随时间所包含的潜在主题，如 Whiting 等[62]指出包含时间驱动主题的查询包含高度可变的子主题，提出从由 Wikipedia 层级结构构建的结构化数据中识别查询中所包含的可能子主题；Zhou 等[63]通过统计用户浏览 Wikipedia 消歧网页次数随时间变化情况分析查询子主题的时间动态性，再利用计算机仿真探讨查询子主题动态对多样化评价的影响。

总体来说，相对时态意图识别研究来说，目前查询动态子主题识别研究比较少。其中，在基于日志方法中，时态意图变化趋势预测研究只能预测周期性变化查询的意图，还缺乏对非周期性变化查询意图的预测研究；在基于文档集方法中，如何能获得有效的能表征查询时态属性的文档集是该研究中关键问题之一。

4 时态感知排序

文档排序是检索系统最核心的部分，在很大程度上决定了检索系统的质量好坏与用户满意度。与一般检索系统排序不同的是，时态信息检索需将文档与查询中的时态信息融合到检索排序模型中。综合已有研究，时态检索排序方法主要分为近因敏感排序与时间依赖性排序两类[16]。

4.1 近因敏感排序

近因敏感排序(Recency-based Ranking)的目的是为查询返回最新文档集，即在主题同等相关的条件下，越新的文档排序越靠前。其研究方法主要有三类：

(1) 融合文档新颖性的排序模型。现有的代表性工作是将时间信息作为文档先验概率融入统计语言检索模型。作为近因排序算法的最早研究者，Li 等[64]在扩展一般语言模型[65]基础之上提出了时间语言模型，即在考虑文档先验概率 P(d)时，不同创建时间文档的权值 P(d|Td)不同，越是最新文档其权值越高；Efron 等[66]扩展了 Li 等[64]的研究工作，认为指数分布参数在不同查询背景下取值不一样，故提出了基于查询的语言模型，且该模型在 TERC 和微博数据上取得了更优的结果；Jatowt 等[65]假设被频繁更新的文档更有可能包含新颖内容，故在主题相关性一致的情况下，被频繁更新或更新幅度较大的文档更有可能排名靠前；Elsas 等[67]为探讨文档动态性与相关性排序之间的关系，先根据

词的时间属性对其加权，再利用语言模型进行文档排序，其实验结果表明该方法有助于导航类检索性能的提升; Aji 等[68]提出一种新的词加权模型即校正历史分析(Revision History Analysis, RHA)模型，在该模型中查询词权值与该词出现在文档不同版本中次数相关，且设定该词出现在较老版本中权值高于出现在较新版本中的权值，然后将 RHA 模型应用到 BM25 与生成统计语言模型中对文档进行排序; Nguyen 等[55]在已有查询多样化排序的基础上提升最新文档的权重，以此实现近因敏感的多样化排序; Daiz[69]通过融合一些新闻网页内容实现近因敏感检索问题。

(2) 基于网络中文档链接结构的排序模型。Berberich 等[70]基于链接分析，提出 T-Light 与 T-Rank 两种排序方法，这两种方法均利用网页的新颖度(即最近更新文档的时间标识)与更新频率来检索最新文档; Cho 等[71]为解决 PageRank 算法中无法提高新创建网页权值的问题，通过分析网络链接结构与分析其结构演化情况提出一种新的排序方法; Li 等[72]尝试根据文档最新时间，为 PageRank 设置非固定的阻尼因子; Zhang 等[73]提出若文档的标题、URL 以及锚文本中出现了最新时态特征，则这些文档应该赋予更高的权值; Dai 等[74]根据网页随时间的变化性以及被链接网页的新颖性来衡量网页的权威性，并将这些信息融合到时间排序概率模型中;

(3) 基于机器学习的排序模型。首先通过人工标注查询及其与之相关的文档集(即 query-url 对)，根据查询与文档之间的相关性级别为每个 query-url 标注相关的相关性分数，这些相关性分数将作为排序学习模型最终的分类类别，最后选取分类特征表示每个 query-url 对，训练分类模型预测结果相关性分数: Dong 等[50]首先识别出近因敏感查询，通过选取近因相关特征(如时间标识相关特征、链接相关特征、WebBuzz 相关特征与网分类相关特征)训练分类模型对近因查询的结果进行排序，而对非近因查询采取另外的排序方法; 与以上方法不同的是，Dai 等[75]首先通过伪相关反馈思想为每个查询构建时间伪文档，再根据每个查询时间伪文档信息为文档赋予不同权值，该方法降低了因意图识别的不准确性给最终实验结果造成的影响。

4.2 时间依赖性排序

时间依赖性排序(Time-dependent Ranking)的目的是为查询返回不同时间段的文档，其核心技术是如何将时间段信息融合到排序模型。研究内容主要包括: 融合时态表达式的排序模型、时态多样化检索与特定类型信息中时间信息排序。

(1) 排序模型中融合时态表达式的相关研究有: Arikan 等[76]从 1997 年至 2000 年的 Wikipedia 数据集文档中抽取与查询词相关的时态表达式并将其融合到语言模型中，该方法的核心是计算如何从文档查询表达式中生成查询中时态表达式; Berberich 等[77]分别利用纽约时报(New York Time)的标注语料(1987 年–2007 年)以及 Wikipedia(2009)数据集中的时态表达式，再将其融合到查询似然语言模型中，即查询中的文本和时间部分分别由文档中的文本和时间部分独立地生成; Brucato 等[78]在不借助任何概率模型的情况下，通过计算查询与文档之间关键词相似性来融合查询与文档之间的时间相似性实现时间信息检索; Jin 等[79]通过线性插值三因素(文本相似度、时间相似度和网页重要性)对查询结果进行排序，其中文本相似度主要考虑查询出现在文档集中的频率以及位置，时间相似度表示查询中时间与文档集中时间的交集，网页重要性通过 PageRank 算法计算; Metzler 等[1]从查询日志中识别出查询可能包含的年份限制，通过计算查询与文档中时间相似性实现排序; Kanhabua 等[80]借助纽约时报标注语料(1987 年–2007 年)提出基于学习排序(Learning-to-Rank)技术的时间敏感排序模型，为训练该模型提出了基于时间与实体的两类特征，最终实验结果表明 SVM MAP Learning-to-Rank 模型优于 Berberich 等提出的方法[77]; Chang 等[81]利用从查询日志中获取的用户时间点击信息，根据用户在不同时间段的意图实现对查询结果进行重排序; Costa 等[82]提出时间相关的排序模型，即首先识别出为期 14 年的网页数据集可能涉及到的时间区间，且为每个时间区间构建一系列查询–文档特征向量，再为每个时间区间训练相关排序模型; Alonso 等[83]提出根据查询出现在显式、隐式与相对时间表达式中的频率对某个类簇文档进行排序; Strötgen 等[84]提出利用 BM25 模型融合查询中情景、时间与地理因素在某文档中的距离为文档进行排序; Mishra 等[85]通过计算地理与时间表达式在查询最初返回结果中的次数对查询结果进行重排序。

(2) 时态多样化检索的相关研究有: NTCIR-12 Temporalia 任务中设立了时态多样化检索(Temporal

Diversified Retrieval)子任务，其任务目标是，给定一个查询主题，要求参与者返回与查询主题和四大时态意图(当前、过去、将来与非时间)相关的文档集合，其相关研究有，Gupta 等[86-87]首先利用伪相关文档方法识别查询中可能包含的潜在时间段，再将每个时间片段作为查询结果可能的分面，再借助概率模型实现查询结果多样化；Hou 等[88]根据文档与每个查询时态子主题的相关性分数以及文档中时态表达式与时态意图相关性进行时态多样化排序；Fernando 等[48]采用学习排序方法实现时态多样化检索，其中，所选取的查询–文档特征主要包括：文档中动词时态特征、文档与查询的主题相关度以及文档与每个时态意图类别中时态信息相关性分数特征等。

(3) 特定类型信息中时间信息排序的相关有：Pasca[89]构建了一个时间问答系统；Strotgen 等[84]在排序模型中综合考虑文本、时间与地理查询词在文档之间的距离使得查询结果满足用户时间与地理的需求；在时间图片检索中，Dais 等[90]首先利用 Ephemeral 聚类方法对网络搜索引擎返回结果进行聚类，查询扩展相关时间，再利用扩展后的查询进行图片检索，基于图片视觉特征训练分类模型，对返回的图片进一步筛选出特定时间区间的图片；Kim 等[91]尝试从 Flickr 数据集中抽取图片的时间模式(如图片的拍摄时间)对图片进行排序；Efron[66]在探讨微博检索时，采用查询依赖的语言模型将时间属性融合到文档排序中，利用统计生存分析(Survival Analysis)中局部最大似然估计参数；卫冰洁等[92]在实现时间感知的微博检索中，在假设“越靠近热门时刻，文档越重要”基础上提出基于热门时刻的 4 个系列模型(HTIMs)。

从以上研究可以看出，学界已对时态感知排序模型进行了大量探讨且取得了一定成果，且设立了与时态信息检索相关的评测平台，具体信息如表 1 所示。尽管如此，当前时态信息检索模型还存在着如下问题：实验的可重复性较低，从以上内容可以看出，T-IR 模型测试的主要数据集 Wikipedia 与新闻数据集，因这些数据集具有流动性，故针对在不同时间段的数据集，检索模型中所调节的参数值会存在着差异性；大多数检索模型都假设查询中词是相互独立的，而忽略了查询中词在特定时间段内的依赖性，影响了最终排序的准确度；最终排序结果只能满足查找当前或者某时间段信息的需求，而难以满足用户在更细时态粒度(如某具体日期或具体时刻)中的特定需求。

表 1　与 T-IR 相关的主要评测平台

相关会议名称	会议主要任务	数据集内容	数据集时间跨度	实验结果评价指标
SemEval 2015 时间与空间任务(SemEval 2015 - Time and Space Track)①	与实体相关事件的识别；时态性问答；时态临床信息抽取；空间信息识别等	新闻、论文、维基百科、博客与临床数据集	1960 年–2014 年	F1 值(F1-score)、召回率(Recall)与准确率(Precision)
TREC 时态摘要任务(TREC Temporal Summarization Track)②	提取某事件相关的实时性摘要信息	TREC 知识库扩展数据集(TREC KBA Stream Corpus): 来自于新闻或者其他社交媒体中带有时间戳的文档	2011 年 10 月–2013 年 2 月中旬	(归一化)期望获益指标(nEG(S))、全面性指标(Comprehensiveness Metric, C(S))、期望延迟指标(Excepted Latency Metric, E[latency])及综合以上三类评测指标的归一化期望延迟获益的调和平均值指标(Harmonic Mean of normalized EL, $EG_\tau(S)$)与延迟全面性性指标(Latency Comprehensiveness, C_τ (S))
TERC 知识资源扩展任务(TRCE Knowledge Base Acceleration Track: KBA)③	通过时态排序筛选出与预定义实体相关的文档，并以此来扩展知识资源(如 Wikipedia)	TREC 知识库扩展数据集(TREC KBA Stream Corpus)	2011 年 10 月–2013 年 2 月中旬	F_1 准确度指标(F_1 Accuracy)与 Scaled Utility 指标

①http://alt.qcri.org/semeval2015/index.php?id=tasks.

②http://trec.nist.gov/pubs/call2016.html.

③http://trec-kba.org/.

(续表)

相关会议名称	会议主要任务	数据集内容	数据集时间跨度	实验结果评价指标
NTCIR 时态信息获取任务(NTCIR Temporal Information Access Temporalia)①	时态意图消歧(Temporal Intent Disambiguation: TID); 时态信息检索(Temporal Information Retrieval, TIR) 时态多样化检索(Temporally Diversified Retrieval: TDR)	英文数据集: 由 LivingKnowledge 项目创建的 "LivingKnowledge 新闻和博客标注子数据集"; 中文数据集: Sogou 全网新闻数据集(SogouCA)与 Sogou 互联网语料库(SogouT)	英文数据集: 2011 年 5 月–2013 年 3 月; 中文数据集: SogouCA, 2012 年 6 月–2013 年 7 月; SogouT, 2008 年 11 月	TID 子任务的评测指标: 平均每类别的绝对损失(Averaged Per-class Absolute Lose)与平均余弦相似度(Averaged Cosine Similarity); TIR 子任务的评测指标: P@20、nDCG@20 与 Q@20 指标; TDR 子任务的评测指标: a-nDCG 与 D#nDCG 指标
TREC 微博任务中 Tweet 时间表生成任务(Tweet TimeLine Generation Task of the TREC Microblog Track: TTG)②	返回在时间点 t 之前与查询 Q 相关 Tweet 的摘要信息	TREC 微博数据集(TREC Microblog Dataset)	2014 年	聚类准确率(Cluster Precision))、加权聚类召回率(Weighted Cluster Recall)与非加权聚类召回率(Unweighted Cluster Recall)

5 结 语

本文详细介绍了时态信息检索中文档中时态信息抽取、查询中时态信息识别与时态感知排序等核心问题。从综述中可以看出, 经过多年努力, 时态信息检索取得了较大进展, 但该领域仍存在着如下问题和挑战: 国外对时态检索研究比较多, 而国内的相关研究甚少; 文档关注时间的识别仅停留在对隐式、隐式与相对时态表达式的抽取与排序, 而通过考虑实体和事件信息表达文档来抽取关注事件的研究存在不足; 缺乏对非周期变化查询将来意图预测的相关研究; 根据流动性数据集检索模型构建实验的可重复性实验较低, 且缺乏对 T-IR 检索模型进行评价的统一评测平台。基于此, 时态信息检索未来的可能发展方向包括: 构建标准化的评测数据集, 便于对各检索模型进行有效对比分析; 构建无参数时态信息检索模型, 便于提高检索模型在各实验数据集中的可重复性研究; 时态检索结果的可视化, 便于用户理解信息随时间变化规律, 及时发现所需信息; 实现将来时间信息检索, 以此预测规律或非规律变化事件的未来趋势(如某电影何时会流行、自然灾害何时发生), 有助于决策支持; 时态信息在信息检索其他研究领域的应用, 如时态多样化检索、时态–空间信息检索、时态问答系统以及基于时态信息的检索结果自动摘要与聚类。

参考文献:

[1] Metzler D, Jones R, Peng F, et al. Improving Search Relevance for Implicitly Temporal Queries[C]// Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2009: 700-701.

[2] 孙逸雪. 基于时态信息的主题搜索引擎的研究与实现[D]. 合肥: 中国科学技术大学, 2009. (Sun Yixue. Research and Implementation of a Time-based Focused Search Engine[D]. Hefei: University of Science and Technology of China, 2009.)

[3] 汤庸, 汤娜, 叶小平. 时态信息处理技术研究综述[J]. 中山大学学报: 自然科学版, 2003, 42(4): 4-8. (Tang Yong, Tang Na, Ye Xiaoping. Review on the Technology of Temporal Information Processing [J]. Journal of Sun Yat-Sen University: Natural Science Edition, 2003, 42(4): 4-8.)

[4] 陈磊. 不确定时态信息的粒度建模及其时态关系研究[D]. 广州: 广州工业大学, 2015. (Chen Lei. Research on Granularity Modeling and Temporal Relations of Uncertain Temporal Information [D]. Guangzhou: Guangzhou University of Technology, 2015.)

[5] 舒忠梅, 左亚尧, 张祖传. 时态信息的语义抽取与排序方法研究及系统实现[J]. 计算机工程与科学, 2014, 36(8): 1609-1614. (Shu Zhongmei, Zuo Yarao, Zhang Zuchuan. Study on Extraction and Ranking of Temporal Semantics and System Implementation [J]. Computer Engineering & Science, 2014, 36(8): 1609-1614.)

①https://sites.google.com/site/ntcirtemporalia/.

②https://github.com/lintool/twitter-tools/wiki/TREC-2015-Track-Guidelines.

[6] Alonso O R. Temporal Information Retrieval [M]. University of California at Davis Davis, 2008.

[7] Campos R, Dais G, Jorge A, et al. Survey of Temporal Information Retrieval and Related Applications[J]. ACM Computing Surveys, 2014, 47(2): 1-41.

[8] TimeML Speficication 1.0 [EB/OL]. [2016-07-23]. http://www.timeml.org.

[9] Nunes S, Ribeiro C, David G. Using Neighbors to Date Web Documents[C]//Proceeding of the 9th Annual ACM International Workshop on Web Information and Data Management. 2007: 129-136.

[10] De Jong F, Rode H, Hiemstra D. Temporal Language Models for the Disclosure of Historical Text[C]//Proceedings of the 16th International Conference of the Association for History and Computing. 2005: 161-168.

[11] Kanhabua N, Nørvåg K. Improving Temporal Language Models for Determining Time of Non-time Stamped Documents[C]//Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries. 2008: 358-370.

[12] Chambers N. Labeling Documents with Timestamps: Learning from Their Time Expressions[C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Stroudsburg: Association for Computational Linguistics. 2012: 98-106.

[13] Kotsakos D, Lappas T, Kotzias D, et al. A Burstiness-aware Approach for Document Dating[C]//Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. 2014: 1003-1006.

[14] Garcia-Fernandez A, Ligozat A L, Dinarelli M, et al. When Was it Written? Automatically Determining Publication Dates[C]//Proceedings of the 18th International Conference on String Processing and Information Retrieval. 2008: 221-236.

[15] Tilahun G, Feuerverger A, Gervers M. Dating Medieval English Charters[J]. The Annals of Applied Statistics, 2012, 6(4): 1615-1640.

[16] Zhao Y, Hauff C. Sub-document Timestamping of Web Documents [C]//Proceedings of the 38th International ACM SIGIR Conference on Research on Development in Information Retrieval. 2015: 1023-1026.

[17] Salah H M, Nelson M L. Carb on Dating the Web: Estimating the Age of Web Resources[C]//Proceedings of the 22nd International Conference on World Wide Web (Companion). 2013: 1075-1082.

[18] Prokhorenkova L O, Prokhorenkov P, Samosvat E, et al. Publication Date Prediction Through Reverse Engineering of the Web[C]//Proceedings of the 9th ACM International Conference on Web Search and Data Mining. 2016: 123-132.

[19] Schilder F, Habel C. Temporal Information Extraction for Temporal Question Answering [R/OL]. http://aaaipress.org/Papers/Symposia/Spring/2003/SS-03-07/SS03-07-006.pdf.

[20] Mani I, Wilson G. Robust Temporal Processing of News[C]//Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. 2000: 69-76.

[21] Strötgen J, Gertz M. HeidelTime: High Quality Rule-based Extraction and Normalization of˙ Temporal Expressions [C]//Proceedings of the 5th International Workshop on Semantic Evaluation. 2010: 321-324.

[22] Chang A, Manning C. SUTIME: A Library for Recognizing and Normalizing Time Expressions [EB/OL]. [2016-07-26]. http://www-nlp.stanford.edu/pubs/lrec2012-sutime.pdf.

[23] Strötgen J, Alonso O, Gertz M. Identification of Top Relevant Temporal Expressions in Documents[C]// Proceedings of the 2nd Temporal Web Analytics Workshop. 2012: 33-40

[24] Jatowt A, Kawai K, Tanaka K. Detecting Age of Page Content[C]//Proceedings of the 9th Annual ACM International Workshop on Web Information and Data Management. 2007: 137-144.

[25] Jatowt A, Yeung C M A, Tanaka K. Estimating Document Focus Time[C]//Proceedings of the 22nd ACM International Conference on Information & Knowledge Management.2013: 2273-2278.

[26] Jatowt A, Ching M, Au Y, et al. Generic Method for Detecting Focus Time of Documents [J]. Information Processing & Management, 2015, 51(6) : 851-868.

[27] Zhao X, Jin P, Yue L. Discovering Topic Time from Web News [J]. Information Processing & Management, 2015, 5(6): 869-890.

[28] Kumar A, Baldridge J, Lease M, et al. Dating Texts Without Explicit Temporal Cues [J]. arXiv Preprint. arXiv: 1211.2290, 2012.

[29] Spitz A, Strötgen J, Bogel T. Terms in Time and Times in Context: A Graph-based Term-Time Ranking Model[C]//Proceedings of the 24th International Conference on World Wide Web. 2015: 1375-1380.

[30] Hoffart J, Suchanek F M, Berberich K, et al. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia [J]. Artificial Intelligence, 2013, 194: 28-61.

[31] Kanhabua N, Nørvåg K. Determining Time of Queries for Reranking Search Results[C]//Proceedings of the 14th European conference on Research and Advanced Technology

for Digital Libraries. 2010: 261-272.

[32] Georgescu M, Kanhabua N, Krause D, et al. Extracting Event-related Information from Article Updates in Wikipedia [C]//Proceedings of the 35th European Conference on Advances in Information Retrieval Heidelberg: Springer-Verlag Berlin. 2013: 254-266.

[33] Ciglan M, Nørvåg K. WikiPop: Personalized Event Detection System Based on Wikipedia Page View Statistics[C]//Proceedings of 19th ACM International Conference on Information and Knowledge Management. 2010: 1931-1932.

[34] 宋巍. 基于主题的查询意图识别研究[D]. 哈尔滨: 哈尔滨工业大学, 2013. (Song Wei. Research on Topic Based Query Intent Identification [D]. Harbin: Harbin University of Science and Technology, 2013.)

[35] Vlachos M, Meek C, Vagena Z, et al. Identifying Similarities, Periodicities and Bursts for Online Search Queries[C]//Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data. 2004: 131-142.

[36] Parikh N, Sundaresan N. Scalable and Near Real-time Burst Detection from eCommerce Queries[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008: 972-980.

[37] Kulkarni A, Teevan J, Svore K M, et al. Understanding Temporal Query Dynamics[C]//Proceedings of the 4th International Conference on Web Search and Web Data Mining. 2010: 167-176.

[38] Zhang R, Konda Y, Dong A, et al. Learning Recurrent Event Queries for Web Search[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Proceeding. 2010: 1129-1139.

[39] König A C, Gamon M, Wu Q. Click-through Prediction for News Queries[C]//Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2009: 347-354.

[40] Ren P, Chen Z, Ma J, et al. Detecting Temporal Patterns of User Queries [J]. Journal of the Association for Information Science and Technology, 2015, 68(1): 113-128.

[41] Jones R, Diaz F. Temporal Profiles of Queries[J]. ACM Transactions on Information Systems, 2007, 25(3): 1-31.

[42] Campos R, J aorge A, Dias G. Using Web Snippets and Query-logs to Measure Implicit Temporal Intents in Queries [C]//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2011.

[43] Hideo J, Jatowt A, Blanco R. Overview of NTCIR-11 Temporal Information Access (Temporalia) Task[C]//Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies. 2014.

[44] Hideo J, Jatowt A, Blanco R, et al. Overview of NTCIR-12 Temporal Information Access (Temporalia-2) Task[C]//Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies. 2016.

[45] Yu H, Kang X, Ren F. TUTA1 at the NTCIR-11 Temporalia Task[C]//Proceedings of the 11the NTCIR Conference on Evaluation of Information Access Technologies. 2014.

[46] Zhao Y, Hauff C. Temporal Query Intent Disambiguation Using Time-Series Data[C]//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2016: 1017-1020.

[47] Pei J, Huang D, Ma J, et al. DUT-NLP-CH@ NTCIR-12 Temporalia Temporal Intent Disambiguation Subtask[C]//Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies. 2016.

[48] Fernando Z T, Jaspreet S, Avishek A. L3S at the NTCIR-12 Temporal Information Access (Temporalia-2) Task[C]//Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies. 2016.

[49] Amodeo G, Blanco R, Brefeld U. Hybrid Models for Future Event Prediction[C]//Proceedings of the 20th ACM International Conference on Information and Knowledge Management. ACM, 2011: 1981-1984.

[50] Dong A, Chang Y, Zheng Z, et al. Towards Recency Ranking in Web Search[C]//Proceedings of the 2nd ACM International Conference on Web Search and Data Mining. ACM, 2010: 11-20.

[51] Styskin A, Romanenko F, Vorobyev F, et al. Recency Ranking by Diversification of Result Set[C]//Proceedings of the 20th ACM International Conference on Information and Knowledge Management. ACM, 2011: 1949-1952.

[52] Cheng S, Arvanitis A, Hristidis V. How Fresh Do You Want Your Search Results? [C]//Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. ACM, 2013: 1271-1280.

[53] Kanhabua N, Nguyen T N, Nejdl W. Learning to Detect Event-Related Queries for Web Search[C]//Proceedings of the 24th International Conference on World Wide Web. ACM, 2015: 1339-1344.

[54] Nguyen T N, Kanhabua N, Nejdl W, et al. Mining Relevant Time for Query Subtopics in Web Archives[C]//Proceedings of the 24th International Conference on World Wide Web.

ACM, 2015: 1357-1362.

[55] Nguyen T N, Kanhabua N. Leveraging Dynamic Query Subtopics for Time-aware Search Result Diversification [C]// Proceedings of the 36th European Conference on Advances in Information Retrieval. Switzerland.Springer, 2014: 222-234.

[56] Shokouhi M, Radinsky K. Time-sensitive Query Auto-completion[C]// Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2012: 601-610.

[57] Shokouhi M. Detecting Seasonal Queries by Time-series Analysis[C]// Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2011: 1171-1172.

[58] Radinsky K, Svore K, Dumais S, et al. Modeling and Predicting Behavioral Dynamics on the Web[C]// Proceedings of the 21st International Conference on World Wide Web. ACM, 2012: 599-608.

[59] Gupta D, Berberich K. Identifying Time Intervals of Interest to Queries[C]//Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. ACM, 2014: 1835-1838.

[60] Gupta D, Berberich K. Temporal Query Classification at Different Granularities[C]// Proceedings of the 22nd International Symposium on String Processing and Information Retrieval(SPIRE 2015). 2015: 157-164.

[61] Dakka W, Gravano L, Ipeirotis P G. Answering General Time Sensitive Queries [J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(2): 220-235.

[62] Whiting S, Zhou K, Jose J, et al. Temporal Variance of Intents in Multi-faceted Event-driven Information Needs[C]// Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2013: 989-992.

[63] Zhou K, Whiting S, Jose J M, et al. The Impact of Temporal Intent Variability on Diversity Evaluation[C]// Proceedings of the 35th European Conference on Advances in Information Retrieval. Heidelberg. Springer-Verlag, 2013: 820-823.

[64] Li X, Croft W B. Time-based Language Models[C]// Proceedings of the 12th International Conference on Information and Knowledge Management. ACM, 2003: 469-475.

[65] Jatowt A, Kawai Y, Tanaka K. Temporal Ranking of Search Engine Results[C]//Proceedings of the 6th International Conference on Web Information Systems Engineering. Heidelberg. Springer-Verlag , 2005: 43-52.

[66] Efron M. Query-specific Recency Ranking: Survival Analysis for Improved Microblog Retrieval[C]// Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2012.

[67] Elsas J L, Dumais S T. Leveraging Temporal Dynamics of Document Content in Relevance Ranking[C]// Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. ACM, 2010: 1-10.

[68] Aji A, Wang Y, Agichtein E, et al. Using the Past to Score the Present: Extending Term Weighting Models Through Revision History Analysis[C]//Proceedings of the 19th ACM International Conference on Information and Knowledge Management. ACM, 2010: 629-638.

[69] Diaz F. Integration of News Content into Web Results[C]// Proceedings of the 2nd ACM International Conference on Web Search and Data Mining. ACM, 2009: 182-191.

[70] Berberich K, Vazirgiannis M, Weikum G. Time-aware Authority Ranking [J]. Internet Mathematics, 2005, 2(3): 301-332.

[71] Cho J, Garcia-Molina H. Estimating Frequency of Change [J]. ACM Transactions on Internet Technology, 2005, 3(3): 256-290.

[72] Li X, Liu B, Yu P. Time Sensitive Ranking with Application to Publication Search[A]// Link Mining: Models, Algorithms, and Applications[M]. Springer New York, 2010.

[73] Zhang R, Chang Y, Zheng Z, et al. Search Result Re-ranking by Feedback Control Adjustment for Time-sensitive Query [C]// Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2009.

[74] Dai N, Davison B. Freshness Matters: In Flowers, Food, and Web Authority[C]//Proceedings of 33rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2010: 114-121.

[75] Dai N, Shokouhi M, Davison B D. Learning to Rank for Freshness and Relevance[C]//Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2011: 95-104.

[76] Arikan I, Bedathur S, Berberich K. Time Will Tell: Leveraging Temporal Expressions in Information Retrieval [C]// Proceedings of the 2nd ACM International Conference on Web Search and Data Mining. ACM, 2009.

[77] Berberich K, Bedathur S, Alonso O, et al. A Language Modeling Approach for Temporal Information Needs[C]// Proceedings of the 32nd European Conference on Advances

in Information Retrieval. Heidelberg. Springer-Verlag, 2010: 13-25.

[78] Brucato M, Montesi D. Metric Spaces for Temporal Information Retrieval [C]//Proceedings of 36th European Conference on Information Retrieval. Heidelberg. Springer-Verlag, 2014: 385-397.

[79] Jin P, Lian J, Zhao X, et al. TISE: A Temporal Search Engine for Web Contents[C]// Intelligent Information Technology Application, 2008, 3: 220-224.

[80] Kanhabua N, Nørvåg K. Learning to Rank Search Results for Time-Sensitive Queries[C]//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. ACM, 2012: 2463-2466.

[81] Chang P T, Huang Y C, Yang C L, et al. Learning-based Time-sensitive Reranking for Web Search[C]// Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2012: 1101-1102.

[82] Costa M, Couto F, Silva M. Learning Temporal-dependent Ranking Models[C]//Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. ACM, 2012: 757-766.

[83] Alonso O, Gertz M, Baeza-Yates R A. Clustering and Exploring Search Results Using Timeline Constructions[C]// Proceedings of the 18th ACM Conference on Information and Knowledge Management. ACM, 2009: 97-106.

[84] Strötgen J, Gertz M. Proximity 2 -Aware Ranking for Textual, Temporal, and Geographic Queries[C]// Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management. ACM, 2013: 739-44.

[85] Mishra A, Milchevski D, Berberich K. Vocabulary-based Re-ranking for Geographic and Temporal Searching at NTCIR Geotime Task[C]//Proceedings of the 6th NTCIR Conference on Evaluation of Information Access Technologies. 2010: 181-184.

[86] Gupta D, Berberich K. Diversifying Search Results Using Time[C]//Proceedings of the 2016 European Conference on Information Retrieval. 2016: 789-795.

[87] Gupta D, Berberich K. A Probabilistic Framework for Time-Sensitive Search [C] //Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies. 2016.

[88] Hou Y, Xu J, Wang X, et al. HITSZ-ICRC at NTCIR-12 Temporal Information Access Task[C]//Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies. 2016.

[89] Pasca M. Towards Temporal Web Search[C]// Proceedings of the 2008 ACM Symposium on Applied Computing. ACM, 2008: 1117-1121.

[90] Dias G, Moreno J G, Jatowt A, et al. Temporal Web Image Retrieval[C]// Proceedings of the 19th International Symposium on String Processing and Information Retrieval (SPIRE 2012). Heidelberg. Springer-Verlag, 2012: 199-204.

[91] Kim G, Xing E P. Time-sensitive Web Image Ranking and Retrieval via Dynamic Multi-task Regression[C]// Proceedings of the 6th ACM International Conference on Web Search and Data Mining. ACM, .2013: 163-172.

[92] 卫冰洁, 王斌. 面向微博搜索的时间感知的混合语言模型[J]. 计算机学报, 2014, 37(1): 229-237. (Wei Bingjie, Wang Bin. Time-aware Mixed Language Model for Microblog Search [J]. Chinese Journal of Computers, 2014, 37(1): 229-237.)

作者贡献声明:

张晓娟: 提出研究思路, 文献调研分析, 撰写论文;
韩毅: 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: zhangxiaojuan624@gmail.com。
[1]张晓娟. References.zip. 参考文献.

收稿日期: 2016-08-15
收修改稿日期: 2016-11-06

Reviews on Temporal Information Retrieval

Zhang Xiaojuan　Han Yi
(School of Computer and Information Science, Southwest University, Chongqing 400715, China)

Abstract: [**Objective**] This study aims to summarize the research status of temporal information retrieval (T-IR) and to provide theoretical basis for the study of the relevant scholars to better grasp the T-IR problems. [**Coverage**] We first used Google Scholar to search related literatures by typing the keywords “termporal information retireval” in Chinese and English repectively, without time limit. After getting some related literatures, we further used the retrospective method to get more related literatures. Finally, we get 92 literatures totally. [**Methods**] Based on method of literature survey and methods of inducting and summarizing, a survey of the existing literature on temporal information retrieval was presented from the following three aspects: extracting temporal information from document, identifying temporal information in queries and temporal ranking model. [**Results**] The problems and challenges existing in temporal information retrieval are as follows: little related work existing in China while most of related work existing in foreign countries; lack of methods of data collection and data indexing reflecting dynamic characteristics of real network; ignorance of the important role of the entity and event represent time information when identify the focus time of document; lack of the predicting intent for non-periodic queries and the improvement of reproducibility of temporal information retrieval model experiment to be needed. [**Limitations**] This paper did not review the document crawling, document index and corresponding application of temporal information retrieval. [**Conclusions**] The construction of standardized evaluation datasets and non-parameter temporal information retrieval models will be the future research trends of T-IR.

Keywords: Temporal Information Retrieval　Temporal Information　Temporal Intent　Temporal Ranking

Jisc 研究数据共享服务选择 Preservica 数字保存平台

Preservica 于近日宣布，其数字保存平台已被选为面向英国高等教育机构(Higher Education Institutions, HEIs)的 Jisc 研究数据共享服务(Research Data Shared Service, RDSS)试验阶段框架的一部分。这一新的研究数据共享服务将整合多家内容提供商的内容，允许英国的大学和其他高等教育机构轻松存取数据，以便对其进行出版、发现、安全存储，以及长期保存。该服务的最终目标是确保有价值的研究数据的长期可访问性，使其能够在大学之间得到重复利用和共享。

该项目涉及 17 个试点高等教育机构，有大型的、研究密集型机构，也有小型的专科院所。Preservica 将与 Jisc 和这些试点教育机构一起合作开发一个新的管理系统，旨在减轻机构信息技术人员和采购人员的负担。除此次合作之外, Preservica 也向美国的几所大学(包括耶鲁大学)，以及英国的曼彻斯特大学提供数字保存服务。

Preservica 的数字保存平台能有效保护数字信息，确保文件格式不会过时，数字记录可以方便地用于科学研究。Jisc 发起该项目的主要目的之一是为了满足资助者的相关政策，实现研究数据管理的良好实践，汇集有益资源。

“我们很高兴 Preservica 成为我们的研究数据共享服务框架和试点流程的一部分,” Jisc 首席创新官 Rachel Bruce 表示: “Preservica 为我们的项目带来了他们在数字保存方面多年的经验，有助于我们建立一个完全集成的系统。Preservica 已经成功为几所大学提供了数字保存功能，很高兴他们成为我们项目的一部分。”

Jisc 将会采购研究数据管理服务和咨询服务，来支持科研机构的个人研究数据管理要求。该项目的重点是提供一个直观的用户界面，确保研究数据很容易找到，同时，支持机构和外部研究系统之间的互操作性。

(编译自: https://librarytechnology.org/news/pr.pl?id=22118)

(本刊讯)